DOCUMENT RESUME

ED 422 404 TM 028 966

AUTHOR Ruiz-Primo, Maria Araceli; Wiley, Edward; Rosenquist, Anders; Schultz, Susan; Shavelson, Richard J.; Hamilton, Laura; Klein, Steve

TITLE Performance Assessment in the Service of Evaluating Science Education Reform.

SPONS AGENCY National Science Foundation, Arlington, VA.
PUB DATE 1998-04-00
NOTE 30p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 14-16, 1998).
CONTRACT SPA-8751511; TEP-9055443
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Change; *Elementary School Students; Grade 5; *Hands On Science; Instructional Effectiveness; Intermediate Grades; *Performance Based Assessment; Pilot Projects; *Science Education; Test Use
IDENTIFIERS California (San Francisco Bay Area); *Reform Efforts

ABSTRACT

This study explored the sensitivity of a multilevel approach in detecting outcomes of a hands-on science program, exploring whether the instruction of a hands-on unit had any impact on students' performance and whether the estimated magnitude of the impact was different according to the proximity of the assessments to the unit taught. Also studied were whether the impact could be detected at a distal level, and whether the differences observed across types of assessments could be replicated across curricular units. This pilot study was conducted in a medium-sized urban school district in the San Francisco Bay area (California). Five schools, 7 teachers, and 163 fifth graders participated. To implement the multilevel achievement assessment, two performance assessments were selected for each unit, one close and one proximal. Distal assessments included performance assessments developed by the California Systemic Initiatives Assessment Collaborative. Preliminary results suggest that instruction had no impact on student performance. As predicted, close assessments were more sensitive to changes in student performance, while proximal assessments did not show as much impact of instruction. It was not possible to assess the sensitivity of distal assessment because no pretest-posttest data were available. High between-class variation in effect sizes suggests the need to examine the instruction students are receiving. Results were not replicated across the two instructional units. Characteristics of the close and proximal assessments seem to have a higher influence on detecting students' improvement than originally thought. Ongoing studies are discussed. An appendix contains a chart of raw score descriptive statistics. (Contains 19 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 422 404

PERFORMANCE ASSESSMENT IN THE SERVICE OF EVALUATING SCIENCE EDUCATION REFORM*

Maria Araceli Ruiz-Primo, Edward Wiley, Anders Rosenquist,
Susan Schultz, & Richard J. Shavelson
Stanford University

Laura Hamilton & Steve Klein
RAND Corporation

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
M.A. Ruiz-Primo
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

DRAFT

Paper presented at the NCME Annual Meeting
San Diego, CA April 1998

TM028966

*The report herein was supported by the National Science Foundation (No. SPA-8751511 and TEP-9055443). The opinions expressed, however, are solely those of the authors.

PERFORMANCE ASSESSMENT IN THE SERVICE OF EVALUATING SCIENCE EDUCATION REFORM

Maria Araceli Ruiz-Primo, Edward Wiley, Anders Rosenquist, Susan Schultz, & Richard J. Shavelson
Stanford University

Laura Hamilton & Steve Klein
RAND Corporation

As a result of the growing interest in development and implementation of hands-on science curricula, schools are being required to document progress toward national, state, and local educational goals. Funding agencies (e.g., National Science Foundation) want to know whether money spent in implementing *hands-on science curricula* is being well spent, and taxpayers want to know whether they are getting good value for their money. Simply put, "Is science education reform improving student achievement?"

Evaluating the impact of science education reform requires a framework that addresses at least two issues: the meaning of student achievement from the science education reform perspective, and the strategies that can provide information necessary for evaluating that impact. In this paper we focus on the latter issue and briefly touch on the former at the outset. Then we propose an approach for evaluating the impact of hands-on science curricula and provide evidence on the sensitivity of the approach to ascertaining outcomes of a hands-on science program. More specifically, we report on an exploratory study using a *multilevel achievement assessment* of two instructional units from the Full Option Science System (FOSS, 1993) curriculum.

Defining Science Achievement

In recent years, the science education community has focused on defining standards that can be used as a guide for developing and implementing curricula that are more aligned with the purpose of science education reform. The same community has reached consensus that the purpose of school science is to achieve scientific literacy (e.g., Bybee, 1996). The strength of this notion is its acceptability among science educators; however, its ambiguity is evident when translating its meaning into practice. What scientific literacy

means for a fifth grade teacher, a curriculum developer, or an educational researcher may be very different.

Scientific literacy has been thought to have different dimensions. According to Bybee (1996), it involves knowing the concepts of a discipline (vocabulary), relating those concepts in schemes according to the structure of that discipline (conceptual scientific literacy) and applying and using that conceptual literacy to solve problems and discover new information (procedural scientific literacy). These dimensions correspond to some dimensions we have used to define science achievement (Shavelson & Ruiz-Primo, in press): *declarative* knowledge--knowing something; *procedural* knowledge--knowing how to do something; and *strategic* knowledge--knowing which, when, and why specific knowledge is applicable. We prefer the term science achievement rather than scientific literacy for two reasons: (1) achievement is always related to educational experiences and carries the connotation of accomplishment (e.g., Cronbach, 1990; Linn, 1992); and (2) achievement is more widely used to refer to what students know and can do (e.g., Glaser & Linn, 1997).

Adopting a broader definition of achievement inevitably leads to a wider array of measuring instruments than those typically used in achievement testing (see Shavelson & Ruiz-Primo, in press). To obtain a more accurate profile of what students know and can do, different sources of information need to be used. Accordingly, we think that to evaluate the impact of science reform, student achievement should be measured using different instruments that can tap the different dimensions of achievement.

Whether this is possible at the moment is another issue. First, we need to find and develop those instruments. Although new forms of assessments are being explored, there is still a long way to go before we know which assessments are suitable for assessing achievement in a large-scale context. Second, with the low expectations and high criticism put on multiple-choice tests, we need to provide evidence to science educators, science teachers, and the public about the suitability of multiple-choice for tapping some unique important aspects of science achievement.

In this exploratory study we use only performance assessments. One reason is that performance assessments are seen as a good match between

hands-on curricula and the knowledge and skills that need to be assessed according to the goals pursued by the science reform. They are consistent with the way students are supposed to be taught science--doing as well as knowing. Another reason is merely practical. This pilot study focused on the suitability of the model rather than the development of the first aspect of the framework (defining science achievement and methods for measuring it). Other forms of assessments, such as multiple-choice tests, should be considered in future studies.

Evaluating the Impact of Science Reform: A Multilevel Achievement Assessment Approach

Although policy documents at the national (e.g., Benchmarks for Science Literacy/AAAS, 1993; National Science Education Standards/NRC, 1996) and state (e.g., State Science Frameworks) level have been designed to guide the translation of the science reform purpose into specific programs and practices, their translation into practice has been difficult (e.g. Bybee, 1996). As a result, students may not be learning the same things across schools, districts, and states. This situation poses serious challenges to evaluating the impact of the implementation of the new curricula.

On the one hand, it has been argued that the statewide assessments students take may not be directly tied to the curriculum they are studying. In other words, students are tested, but not necessarily on what they have been learning in their classroom (e.g., American Federation of Teachers, 1997) or with the measurement methods that match the way they have been taught (e.g., Dowling, 1987; Hein, 1987). On the other hand, administrators and policy makers are more interested in the general level of knowledge and competencies that students have gained from their science instruction. Therefore, statewide/nationwide assessments avoid, by design, special topics of concentration on specific subject matter taught to only a fraction of the students being tested. This situation sets up a tension between the knowledge and competencies students are able to demonstrate on a particular assessment and those they may have which the test does not in fact probe (e.g., Raizen, Baron, Champagne, Haertel, Mullis, & Oakes, 1990).

To solve this tension we propose a *multilevel achievement assessment*. We think that if science reform is having an impact on student's achievement,

this impact should be located at different levels, first at the local classroom curriculum level, and then, hopefully, in transfer to higher cross-school levels as measured by statewide assessments. Evaluating students with achievement measures at different distances from the science curriculum they study provides a better picture of the extent of the effect that science instruction is having (regardless of the specific curriculum being taught) than using distal mesures alone. A multilevel achievement assessment, then, estimates the impact of a hands-on science curriculum at different distances.

The idea of multilevel achievement assessment is based on the Brunswikian "regional reference" approach (see Snow, 1968). This approach classifies variables according to their remoteness from the central process of a subject. Variables are seen as laying in regions or layers increasingly peripheral to a subject. In the context of science reform evaluation, the regional reference approach is used to classify assessments according to their proximity (or remoteness) to student learning of the curriculum implemented.

With multilevel achievement approach, evidence is collected at different distances from the enactment of a curriculum: close, proximal, distal, and remote (Figure 1). At the *close* level, assessments should be curriculum sensitive; they are close to the content and activities of the curriculum. At a *proximal* level, assessments should be designed considering the knowledge and skills relevant to the curriculum, but content (e.g, topics) can be different to the one studied in the unit. At a *distal* level, assessments may be based on state/national standards on a particular domain. At a *remote* level, general measures of achievement should be used.

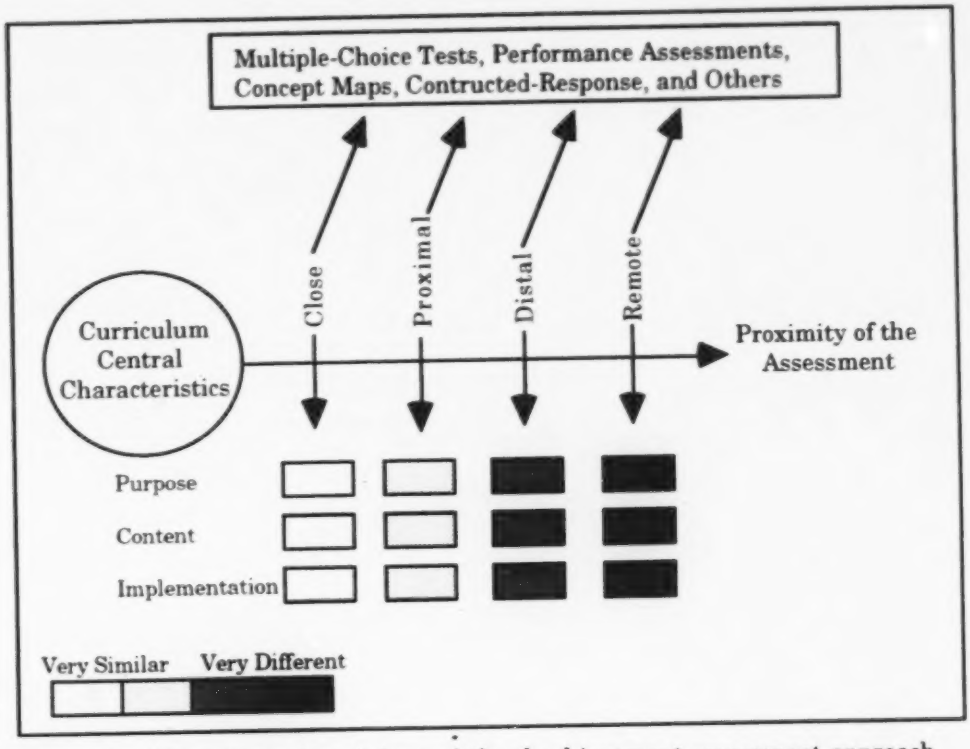

Figure 1. Characterization of the multilevel achievement assessment approach. Classification of assessments according to their proximity to the characteristics of a curriculum.

To establish the proximity of the assessments to the central characteristics of the unit/curriculum, we propose three categories: Purpose, Content, and Implementation (Figure 1, Table 1). These categories do not attempt to be exhaustive, they only provide a guide to capture how different the assessments are to a curriculum. Defining the proximity on each category helps to define a proximity profile for each assessment, like the hypothetical profile in Figure 1.

We think that different proximity profiles under the same proximity classification are likely to occur. It is possible that different proximity profiles within the same category of proximity create differences in the difficulty of assessment--something that can only be studied empirically.

Table 1

Categories and Aspects Used to Define the Proximity of an Assessment

Areas	Aspects
Purpose	• What is the assessment purpose based on? ☐ instructional activity ☐ unit goals ■ curriculum goals ■ national/state standards
Content	• What is the assessment task content based on? ☐ same content domain, topic, concepts and principles of the unit ☐ same content domain and topic, but different concepts of the unit ■ same content domain, but different topic ■ different content domain
Implementation	<u>Assessment task</u> • What is the assessment task based on? ☐ instructional activities implemented in the unit, same problem and procedures to solve it ☐ instructional activities implemented in the unit, problem is different, but procedure to solve is the same ■ an activity not used in the unit, but using same procedures as those used in the unit ■ an activity not used in the unit and procedures used are not the same as those in the unit <u>Assessment task degree of structure</u> • Is the level of directedness/structuredness of the assessment task the same as the instructional activities in the unit? ☐ instructional activities and assessment task have the same level of directedness/structuredness--procedures and data analysis and interpretation are either provided or not provided ☐ instructional activities are low directed/structured and assessment task is highly directed/structured--procedures and data analysis and interpretation are not provided in the instructional activity, but they are in the assessment task ■ instructional activities are highly directed/structured and assessment task is low directed/structured--procedures and data analysis and interpretation are provided in the instructional activity, but they are not in the assessment task ■ instructional activities and assessment task are different in what is provided or not provided

Table 1

Continue

Areas	Aspects
Implementation	<u>Assessment Materials</u> • How similar are the materials used in the assessment task compared to the ones used in the unit? ☐ same as those used in the instructional activities ☐ same as those used in the unit, but in a different instructional activity ■ different to those used in the unit, but comparable in terms of function and purposes ■ different to those used in the unit and not comparable <u>Assessment Measurement Methods</u> • How similar are the measurement methods used in the assessment task to those used in the unit? ☐ measurement focuses on same variables and uses same measuring instruments and procedures ☐ measurement focuses on same variables, and uses same measurement instruments, but different procedures ▒ measurement focuses on same variables and uses different measuring instruments and different procedures ■ measurement focuses on different variables, different measuring instruments, and different procedures

The multilevel achievement assessment also proposes the collection of at least three classes of information (see Wolf, 1990): (1) initial status of the students--who they are and how proficient they are with regard to what they are supposed to learn; (2) students' performance after a period of instruction/ implementation of the unit/program/curriculum--how proficient students are after the period of instruction; (3) implementation of the unit/program/ curriculum--at the very least, one needs to know whether the unit/program/ curriculum was actually implemented, and if so, to what extent.

To provide an accurate picture of the impact of the science education reform, data need to be collected over a period of years considering a pre-post design at all levels of evaluation. Analysis of trends in such longitudinal data will allow comparisons in the absence of pre-post measurements in the future.

The pilot study reported here focused only on some aspects of the approach. We only administered a close and proximal assessment using a pre-post design. The distal assessment was administered only after instruction took place because it had to be administered in conjunction with regularly scheduled district testing.

In this study we focused on the following questions: Does instruction of a hands-on unit have any impact on students' performance? If so, is the estimated magnitude of this impact different depending on whether a close or a proximal assessment is used? If impact is detected by the proximal assessment, is it also captured by the distal assessment? Finally, if differences are observed across types of assessments, are these differences replicable across curricular units?

Method

Subjects. A medium sized urban school district in the Bay Area, which has received NSF support since 1990 to implement hands-on science curricula, participated in the study. Five schools from the 75 elementary schools in the district participated with seven classes/teachers and 163 fifth-graders. One of the units, Variables, was taught in 3 classes (70 students) and the other unit, Mixtures and Solutions, in four (93 students).

Curriculum. FOSS (1993) was the hands-on science curriculum implemented in the school district where the study was carried out. FOSS goals are scientific literacy for all students and instructional efficiency for all teachers. The FOSS curriculum is organized in modules classified according to: (1) content domains--life science, physical science, earth science, and scientific reasoning and technology, and (2) grade levels--kindergarten, Grade 1 and 2, Grade 3 and 4, Grade 5 and 6. Curriculum modules include three main components: a teacher guide, an equipment kit, and a teacher preparation video. Each module has different activities (i.e., sections). Activities are designed in a way that they can be implemented independently.

Two modules, called here units, were selected to implement the multilevel achievement assessment approach: "Variables" and "Mixtures and Solutions." Both units were developed for fifth- and sixth-graders. Variables is one of two units for scientific reasoning and technology and Mixtures and Solutions is one of two units for physical science.

Instrumentation. To implement the multilevel achievement assessment we selected performance assessments because they are considered to provide a closer match between hands-on curriculum and the knowledge and skills that need to be assessed. We selected two performance assessments for each unit, one close and one proximal. The distal assessments included performance assessments developed by the California Systemic Initiatives Assessment Collaborative (CSIAC).

To provide an idea of what *close, proximal*, and *distal* assessments are, we describe one of the units and the three most proximal assessments used to evaluate the impact of instruction (Table 2). In the Variables unit (FOSS, 1993), students are expected to design and conduct experiments; describe the relationship between variables discovered through experimentation; record, graph and interpret data; and use these data to make predictions. During the unit, students identify and control variables, and conduct experiments using four multivariable systems (e.g., Swingers and Lifeboats).

The *close assessment* used to evaluate the Variables Unit was a modified version of the Pendulum Assessment (Stecher & Klein, 1995). In this assessment, students are asked to identify the variable that affects the time it takes a pendulum to complete 10 cycles. Students explore the relationship between the length of a string, the weight of the suspended object, and the periodicity of a pendulum. The scoring system focuses on the correctness of the variable identified, the accuracy of the students' measurements, and their interpretation of the data. The assessment task can be considered as an exchangeable instructional activity with the "Swingers" activity--the procedure for testing the variables is the same used in the instructional activity (see Table 2). Differences between the instructional and the assessment tasks are: (1) the material used to construct the pendulum and to manipulate the suspended weight, and (2) the way the dependent variable is measured (Figure 2).

The *proximal assessment* was the Bottles Assessment (Solano-Flores, Shavelson, Ruiz-Primo, Schultz, & Wiley, 1997; Solano-Flores, & Shavelson, 1997). In this assessment students are asked to explain what makes bottles float or sink. Students are provided with 12 bottles, which vary in size, weight, and color; a tub filled with water, and one tray. Students need to identify the relevant variables that make a bottle float or sink and then explain the relationship between the relevant variables selected and floating or sinking.

Table 2

Description of the Variables Units and the Close, Proximal and Distal Assessments

Dimensions	Variables Unit	Pendulum Assessment	Bottles Assessment	CSIAC Assessments
Purpose	FOSS Expects students to: • Gain experience with the concept of variable and system. • Design and conduct controlled experiments. • Use data to make predictions. • Record and graph data to discover relationships. • Acquire the vocabulary associated with experimentation.	Assessment Task: • Explore the relationship between the length of a string, the weight of the suspended object, and the periodicity of a pendulum.	Assessment Task: • Explore the relationship between volume, weight, and sinking and floating. Bottles provided to students are sinkers or floaters.	Assessment Tasks 1: • Explore properties of objects and materials Assessment Tasks 2: • Explore the magnetic effect of electricity and the effect of magnets.
Content	Content Domain: • Scientific Reasoning and Technology Topic: • Relationships and interactions, cause-and-effect events Concepts: • Concept of variable and relationship between variables	Content Domain: • Scientific Reasoning and Technology Topic: • Relationships between variables Concepts: • Concept of variable and relationship between variables	Content Domain: • Scientific Reasoning and Technology Topic: • Relationships between variables Concepts: • Concept of variable and relationship between variables	Content Domain: • Physical Science Topic: • Properties of objects and materials • Light, heat, electricity, and magnetism Concepts: • Objects' observable properties • Electricity in circuits can produce magnetic effects. • Magnets attract and repel each other and certain kinds of other materials.

Table 2

Continue

Dimensions	Variables Unit	Pendulum Assessment	Bottles Assessment	CSIAC Assessments
Implementation	Activities • Students identify and control variables, and conduct experiments using multivariate systems. (e.g., Swingers and Lifeboats) • Students explore the effect of length and weight on the pendulum behavior • Explore height of boats (cups) and the arrangement of passengers (pennies) as variables for making the boats float. • Use graphs to predict results. • Group-based activities. Directedness Highly structured hands-on activities. Students are provided with: • The variables and the values of the variables to be explored. • The design of the experiment. • The procedures to control the variables. • The procedures to graph the results. Materials Swingers: • Strings, papers clips, meter tape, masking tape, pencils to construct the pendulum. Lifeboats: • Tub with water, paper cups, graduated beakers, pennies, 2 books, scissors. Measurement Method: Swingers: • Frequency of swings in 15 seconds. Teacher times the 15 seconds. • Length is measured in centimeters. Lifeboats: • Volume of boats (cups) using a syringe. • Number of passengers (pennies) the boat can support.	Task • Students are asked to identify the variables that affect the time it takes a pendulum to complete 10 cycles. • Individual tasks Directedness Highly structured. Students are provided with: • The variables and the values of the variables. • The design of the experiment. Students: • Use data to make predictions. Materials • Strings, platform and stick with a hook on it. stopwatch, washers. Measurement Method: • Time it takes the pendulum to complete 10 cycles. Students are provided with an individual stopwatch. • Length is measured in inches.	Task • Students are asked to find out which variables (size, weight, color) make bottles float or sink. • Individual tasks Directedness Less structured. Students are provided with: • A chart to organize variables information Students: • Identify the relevant variables for sinking and floating. • Describe how the relevant variables are related. • Use data to make predictions. Materials • Tub filled with water, 12 bottles different sizes, weights and colors, and marked with letters, tray. Measurement Method: • Students do not measure, but classify bottles according to the values of the three variables (size, weight, color).	Task Assessment Tasks 1: • Students are asked to classified items based on their properties. Assessment Tasks 2: • Students are asked to construct and use a magnet. • Individual tasks. Directedness Less structured. Students are provided with: • A chart to record the groups created. Materials • Bag A with trash items (e.g., newspaper, beans, wires) , Bag B with once-living things, Bag C with material to construct a magnet, and Bag D items to be used with the magnet. Measurement Method: Students do not measure, but classify objects according to their characteristics.

The scoring system focuses on the correctness of the identification of the relevant variables and the accuracy of the explanation provided. Differences between the instructional and the assessment tasks are: (1) the materials used in the assessment are totally different; (2) the procedures used to manipulate the variables are different; and (3) the procedure used in the instructional unit to learn about sinkers and floaters is totally different to the procedure used on the assessment task. Still, the assessment requires knowledge about variables, levels of variables, and how to interpret results (Figure 2).

The *distal* assessment for both units was the CSIAC assessment. The CSIAC assessment was developed based on the standards proposed on the National Science Education Standards (NRC, 1996) and the Benchmark for Science Literacy (AAAS, 1993) and supports the learning goals of the different systemic initiatives funded by NSF in two states. Its purpose is to provide essential information and data on student science achievement for schools/districts and/or projects to report program impact to funders. CSIAC assessment includes a 29-item multiple-choice, two performance assessments, and two optional open-ended questions. Table 2 describes the characteristics of the two CSIAC performance assessments, the only form of assessment we used in this study. The instructional and assessment tasks differ in multiple ways: (1) the focus of the assessment task is on a different domain, physical science; (2) none of the topics learned in the unit (e.g., chemical reactions) are part of the assessment tasks; and (3) the materials, measurement methods, and level of directedness/structure of the assessment tasks are different to the ones used to conduct any of the instructional activities. The left side of Figure 2 provides the proximity profiles of the three assessments used to evaluate the impact of the Variables unit, based on the categories of Table 1.

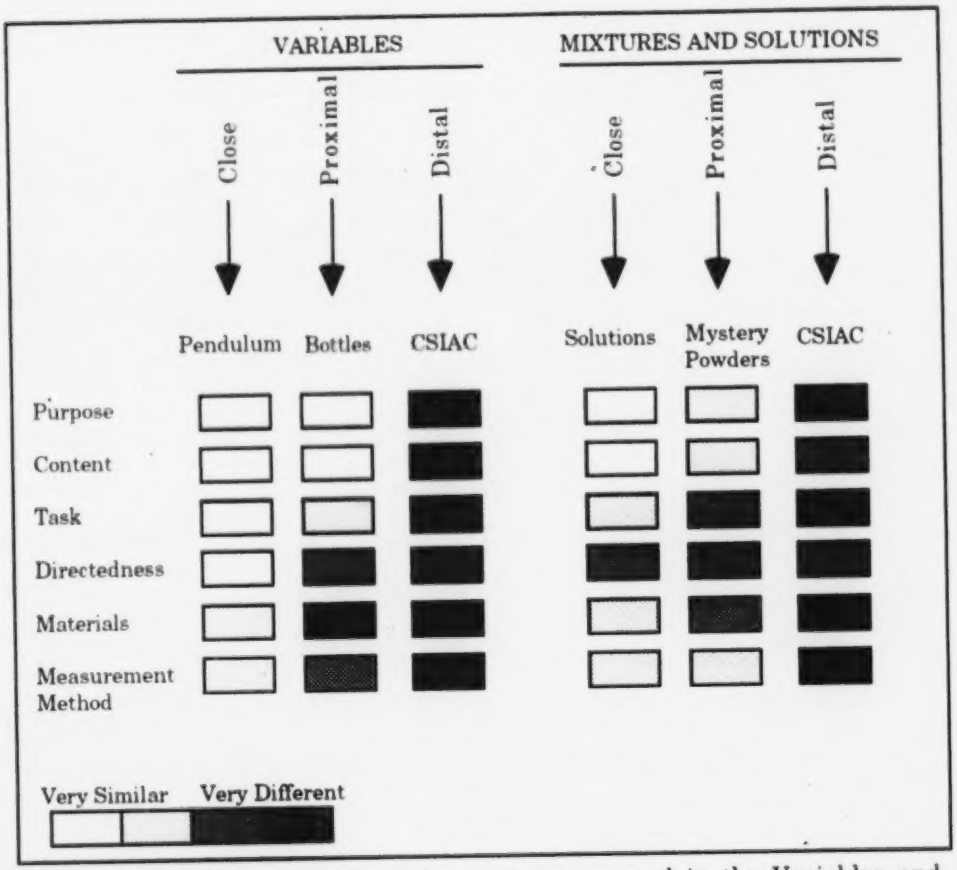

Figure 2. Proximity profiles of the assessments used in the Variables and Mixtures and Solutions units.

Two other performance assessments were used for the Mixtures and Solutions Unit: Solutions--the *close* assessment and Mystery Powders--the *proximal* assessment (Figure 2; Baxter & Shavelson, 1995; Shavelson, Solano-Flores, Ruiz-Primo, in press). The Solution assessment asks students to find out which of three powders is the most and the least soluble in 20 ml. of water. They are asked to provide information about how they conducted the investigation, the results they obtained, and two other questions about solubility (e.g., how they can dissolve the maximum possible powder in a saturated solution). The scoring system focuses on the accuracy of the results and the quality of the procedure used to solve the problem. Differences between the instructional and the assessment tasks are: (1) the material (i.e., the type of powder) used to make the solutions, and (2) the method (i.e., a balance vs the

number of scoops of powder used) used to determine the amount of solid material to saturate the water (Figure 2).

The Mystery Powders assessment has two parts. In Part I students are asked to examine four powders using five tests (sight, touch, water, vinegar and iodine). In Part II, students are asked, based on their observations, to find the content in two mystery powders. In the notebook, students are asked to complete a partially completed 4 x 5 table with their observations for each powder and to provide information about the tests they used to find out the components of powders X and Y. The scoring system focuses on the accuracy of the observations and descriptions, the quality of the evidence provided (confirming, disconfirming, and other), and the correctness of their answers. Differences between the instructional and the assessment tasks are: (1) the material used (i.e., the type of powders), (2) the procedures used to conduct the task (e.g., whereas the instructional activity focuses on solubility/ concentration as properties of matter, the assessment task focuses on chemical reactions) (Figure 2). The CSIAC assessment was considered the distal assessment for this unit. The right side of Figure 2 provides the profile of the assessments.

Procedure. The exploratory study was conducted in two 55-minute sessions during a five week period in the Spring of 1997. Students in each classroom were randomly assigned to one of four sequences of testing according to the type of assessment, close or proximal, taken at pretest and at postest. Sequences were: 1--Close-Close, 2--Proximal-Proximal, 3--Close-Proximal, and 4--Proximal-Close. Although the two first sequences are the main interest of this paper, we wanted to explore the effect that the other two sequences may have on students' performance. Students took the assessments individually before and after studying the Variables and Mixtures and Solutions units. The CSIAC assessment was administered approximately 15 days after the postest in two consecutive days.

Evidence that the units were implemented and completed came from two sources: classrooms observations and students' science journals. Not all classrooms could be observed, however, the collected science journals provide evidence of which activities were completed. We are currently working on ways to evaluate the quality of the students' work as reflected by their journals.

Results

Preliminary analyses have focused on the following questions: Did performance assessments detect an impact of the instruction upon student performance? If such an impact is detected, does the estimated magnitude of the impact depend upon whether a close or proximal assessment was used? If impact is detected at the proximal assessment, is this reflected at the distal level? If differences are observed across the types of assessments, are these differences replicable across curricular units?

If the curriculum has succeeded in effecting change in student science achievement in the specific content area covered by the unit, we would have expected to see an increase in the *close* assessment scores from pretest to postest. Similarly, a pretest-to-postest increase in the proximal assessment scores would indicate that instruction using the FOSS unit brought about achievement in knowledge and skills across different science contents not limited to that specifically covered by the unit. In sum, from the multilevel achievement assessment perspective, we expected the close assessment to be most sensitive to the impact of instruction, assuming that teachers teach the curriculum in a manner consistent with intended practice.

Each student's responses were scored by two raters. Interrater reliability across occasions and assessments was generally high: .89 and .97, on average, across occasions and assessments, for the Variables and Mixtures and Solutions assessments, respectively. To carry out the rest of the analyses, students' scores were averaged across raters.

Impact of Instruction

To determine whether instruction had any impact at all, we focused on differences between the pretest-postest on sequences 1 and 2 (close-close and proximal-proximal). Comparing differences between occasions on the other two sequences was not a straightforward procedure because of differences on scales across assessments (Appendix A shows the means across classes, occasions, sequences, and units).

A series of 2x3 and 2x4 split-plot ANOVAs were carried out to evaluate whether differences between pretest and postest varied across classes and sequences (i.e., sequence 1 and sequence 2). Results were different across units.

For the Variables unit, no significant occasion by class interaction for either sequence ($F_{OxC_{s1}}$=.41, p=.67; $F_{OxC_{s2}}$=.53, p=.60) or main class effect ($F_{C_{s1}}$=1.79, p=.27; $F_{C_{s2}}$=1.75, p=.21) was found, as expected. A significant occasion main effect was found for sequence 1, close-close, ($F_{O_{s1}}$=7.44, p=.02), but not for sequence 2, proximal-proximal ($F_{O_{s2}}$=.02, p=.88).

For the Mixtures and Solutions unit, no significant occasion by class interaction was found on either sequence ($F_{OxC_{s1}}$=1.56, p=.23; $F_{OxC_{s2}}$=.66, p=.59). However a significant class effect was found on both sequences ($F_{O_{s1}}$=3.04, p=.054; $F_{O_{s2}}$=8.17, p=.001). A significant occasion effect was found in sequence 1, close-close ($F_{O_{s1}}$=11.24, p=.003) but not in sequence 2 ($F_{O_{s2}}$=.00, p=1.00), proximal-proximal. A closer look into the class effect indicated that Class 3 created the differences with the rest of the classes. However, for simplicity of presentation we decided to collapse all classes into one and treat all students as one single group, one for the Variables unit, the other one for Mixtures and Solutions.

The repeated-measures ANOVA indicated an impact of instruction only when close assessments were used. It is important to mention that even though there was a significant increase between pretest and postest, low mean scores across the two units using the close assessments show that the knowledge exhibited by the students on the postest was partial and far from the maximum score, specially on the Solutions assessment.

Comparing Close and Proximal Assessments

In the previous section we compared the first two sequences when the same type of assessment, either close or proximal, was used across the two occasions. In this section we compare the four sequences. Acknowledging the fact that no significant difference was detected when proximal assessments were used, did pretest and postest scores vary in the expected direction?

In order to compare the pretest and posttest scores for each of the four assessment sequences, we needed to make score scales comparable across the two types of assessments, close and proximal, within each unit. Since students were randomly assigned to pretest groups, there was no reason to assume that the students taking the close pretest were different in any systematic way from the students completing the proximal pretest. A series of independent-samples *t*-test comparing groups of students taking the same pretest assessment but in different sequences (i.e., sequence 1 and 3; sequence 2 and 4) within each class revealed no

significant differences between groups (p values ranged from .134 to .920). To compare students across sequences within each class we used reading and math scores provided by the school district. Two one-way ANOVAs using sequence as a factor were carried out for each class. No significant differences in reading or math between sequences were found in any class (p values ranged from .10 to .83). It was concluded that students assigned to the different sequences did not differ in any systematic way.

Based on these results, we standardized within class and within task on pretest scores only; we used the pretest means and standard deviations to convert the postest scores.[1] This transformation put everything on a z score metric based on the pretest scores. Because the scores are in the pretest standard deviation metric, the mean postest transformed scores within a given sequence can be considered an effect size.[2]

Table 3

Effect Sizes by Class and Sequence Across the Two Units, Variables and Mixtures

		Units			
		Variables		Mixtures	
Sequence	Class	*n*	Effect Size	*n*	Effect Size
1 Close-Close	1	5	.41	5	1.38
	2	5	.69	5	.43
	3	6	-.07	7	1.28
	All	16	.32	23	1.44
2 Proximal-Proximal	1	7	.01	6	-.33
	2	4	-.25	8	-.24
	3	6	.48	7	.62
	All	17	.12	26	.07
3 Close-Proximal	1	7	.22	6	-.26
	2	5	-.38	7	.42
	3	7	.59	6	-.25
	All	19	.20	24	.37
4 Proximal-Close	1	7	-.16	6	2.71
	2	4	.45	4	-.67
	3	7	1.42	7	1.64
	All	18	.59	20	1.13

[1] $X' = X_{post} - \bar{X}_{pre} / SD_{pre}$

[2] Examination of these effect sizes provides more meaningful information than reliance upon statistical significance of the t-test; the former provides clues to actual pretest-postest change free of sample size limitations, whereas the latter may be heavily influenced by the effects of small sample size.

Trends of effect sizes across sequences varied according to class and unit. This is probably due to differences in the quality of teaching and class composition (i.e., student's characteristics) on the one hand, and to differences in the characteristics of the units and assessments, on the other. To provide a more general and clear picture of the instructional sensitivity of the type of assessment and sequence, we combined classes and calculated the effect sizes (Table 3, Figure 3).

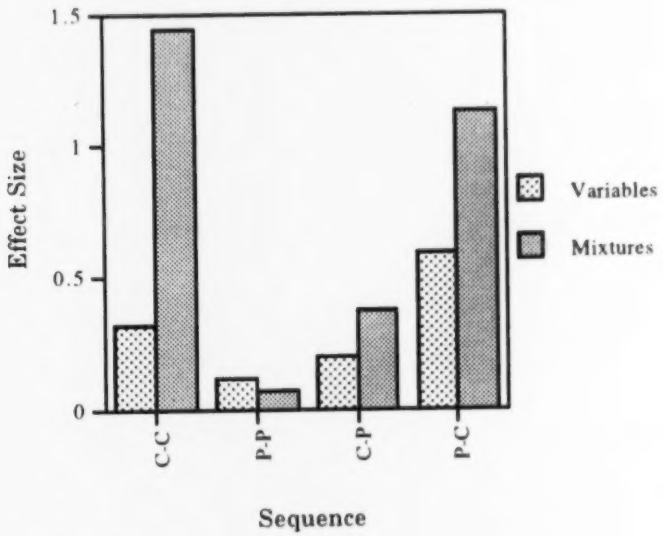

Figure 3. Effect sizes across sequences and units.

Figure 3 shows that patterns across sequences are more similar for both units than first suspected when classes were used as the unit of analysis. It also shows that, overall, instruction had a positive impact. If we focus on the two first sequences, close-close and proximal-proximal, effect sizes indicate that close assessments are more sensitive to detect impact of instruction than proximal assessments. The type of assessment taken as a pretest had an impact on the postest observed scores. When close assessments were taken as a postest (sequence 4), effect sizes are greater than when the proximal assessments were taken as a postest (sequence 3).

Replication of patterns across units, however, is not identical. Differences between the effect sizes for the first two sequences are greatly higher for the Mixtures and Solutions unit, than for the Variables unit. We suspect that differences in the characteristics on the assessments may explain, at least partially, this difference. Profiles of the close and proximal assessments across units are not the same (see Figure 2). For example, the close assessment for the Mixtures and Solutions unit requires more content knowledge for conducting an appropriate investigation than the close assessment for the Variables unit (i.e., whereas the Solutions assessment does not provide procedures to conduct the investigation, Pendulum, a highly structured assessment, provides the exact procedures to conduct the investigation).

Comparing Different Proximity Assessments

Distal assessments measure broader achievement objectives that are necessarily detached from the specific dimensions of an individual unit. Distal assessments are usually administered for comparison of groups of students (e.g., schools, districts) within an educational region. They are typically administered once per year to all students, independently of the instruction students may have encountered up to the point of the assessment.

Because we only have one set of scores on the distal assessment, we could not investigate impact directly. Instead, we used correlations among types of assessments to evaluate the degree to which they rank ordered students similarly. Table 4 provides correlations between the different types of assessments by sequence and across units. For the distal assessment, we provide the correlations with the score obtained on performance assessments (CSIAC-PA).

Comparing Pretest-Postest Correlations. For the Variables unit, the correlations are as might be expected, .76 and .71 when assessments were the same at pretest and posttest, and .54 and .31 when assessments were different. Unfortunately, this pattern was not the same for the Mixtures and Solutions unit (Table 3). The characteristics of the assessments seem to be the reason. The Mystery Powders assessment turned out to be easier for students than the Solutions assessment (means in Appendix A). For example, on sequence 3, students were ranked differently because their scores tended to be low on the pretest (Solutions as pretest) and higher on the postest (Mystery Powders as

postest). On sequence 4, postest scores on Solutions were higher (similar to those observed on postest on sequence 1) and more similar to those obtained on the Mystery Powders assessment. The difference in difficulty may be due in part to the differences of knowledge required to solve the two problems. Solutions requires more content knowledge (e.g., what is saturation and solubility) than the Mystery Powders assessment, where systematic observations and recording are critical skills addressed. Another reason can be the difference in degree of directedness/structuredness of the assessments (Mystery Powders is more structured than Solutions).

Table 4

Correlation Matrix for Scores on Assessments of Different Proximity

		Units			
		Variables		Mixtures	
Sequence	Type of Assessment	Postest	CSIAC-PA	Postest	CSIAC-PA
1 C-C	Pretest	.76**	.64*	.52*	.32
	Postest		.64*		.07
2 P-P	Pretest	.71**	.03	.66*	.43*
	Postest		.20		.57*
3 C-P	Pretest	.54*	.70**	.32	.29
	Postest		.15		.55*
4 P-C	Pretest	.31	.43	.65**	.67*
	Postest		.59*		.63**

** Correlation is significant at the .01 level.
* Correlation is significant at the .05 level.

Comparing Close-Distal and Proximal-Distal Correlations. Correlations between the close and proximal assessments with the distal assessment varied not only across sequences, but also across units. Whereas the highest correlations observed in the Variables unit are those between the close and distal assessments, the highest correlations observed in the Mixtures unit are those between the proximal and distal assessments, independently on whether the close or proximal assessments were administered as pretest and/or postest. Two related factors, assessment difficulty and the degree of structure of the assessment task, seem to influence the correlations (Pendulum and Mystery Powders are the more structured assessment tasks; mean scores for

these assessments are proportionally higher than for the other two assessments; see Appendix A).

Conclusions about sensitivity of distal assessments are difficult to make since patterns of correlations with pretest and posttest scores are similar and varied for close and proximal assessments.

Conclusions

In this study we explored the sensitivity of a multilevel approach in detecting outcomes of a hands-on science program. We examined whether: (1) the instruction of a hands-on unit had any impact on students' performance; (2) the estimated magnitude of the impact was different according to the proximity of the assessments to the unit taught; (3) the impact could be detected at a distal level; and (4) differences observed across types of assessments could be replicable across curricular units.

Our preliminary results led to the following tentative conclusions: (1) Instruction of both units had an impact on students' performance. Significant differences were observed between pretest and postest when close assessments were administered on both occasions (p=.02 and p=.003, for Variables and Mixtures and Solutions respectively). Overall, results were in the predicted direction; close assessments were more sensitive to changes in student performance, whereas proximal assessments did not show as much impact of instruction. (Sensitivity of distal assessments was not possible to evaluate since no pretest-postest data were available.) (2) Whether students can be ranked similarly using distal assessments is still an issue. Correlations found in the preliminary analysis indicate that rank order may depend on the characteristics of the close and proximal assessments. More analyses are being done using the other forms of assessments (i.e., multiple-choice and open-ended scores). We think that a pretest-postest design should be also used for distal assessments to create longitudinal trends that can support future decisions. (3) High between-class variation in effect sizes suggests the need to examine the instruction students are receiving. (4) Results were not replicated across the two instructional units. Differences in the characteristics of the assessments used across units were discussed as a possibility for the differences. (5) Characteristics of the close and proximal assessments seem to have a higher influence on detecting students' improvement than originally

thought. It seems that some of the aspects used to classify assessments (e.g., degree of structure) are more important than others.

We are currently collecting data on 11 schools (20 classes) in the same school district. Some changes were made in the design. Only the two first sequences (close-close and proximal-proximal) are being considered in this new, larger study. Each student is being tested on four occasions, before and after instruction in each of the units. Some changes were also made to the assessments.

The importance of the information gained in this study lies in providing information about the sensitivity of detecting curriculum outcomes according to the proximity of the assessments to the curriculum. Any evaluation of science reform should consider the proximity of the outcomes measures to the curriculum. The use of more distal or remote measures may lead to an erroneous conclusion that the reform has no impact. Nevertheless, if the impact is only evident at the closest possible level, this raises questions about the reform itself.

References

American Federation of Teachers (1997). Making standards matter. Washington, DC: Author.

American Association for the Advancement of Science (1993). Benchmarks for science literacy. Washington, DC: National Academy Press.

Baxter, G.P., & Shavelson, R.J. (1995). Performance assessments in elementary science classrooms: Questions of rater consistency. Unpublished Manuscript.

Bybee, R. W. (1996). The contemporary reform of science education. In J. Rhoton, & P. Bowers (Eds.) Issues in Science Education (pp 1-14). Arlington, VA: National Science Education Leadership Association.

Cronbach, L.J. (1990). Essentials of psychological testing. New York: Harper & Row, Publishers.

Dowling, K.W. (1987). Science achievement testing: Aligning testing method with teaching purpose. In A.B. Champagne & L.E. Hornig (Eds.) This year in school science 1987: Students and science learning (pp 137-152). Washington, DC: American Association for the Advancement of Science.

Full Option Science System (1993). Britannica Science Systems. Chicago, IL: Encyclopaedia Britannica Educational Corporation.

Glaser, R., & Linn, R.L. (1997). Assessment in transition: Monitoring the nation's educational progress. Stanford, CA: The National Academy of Education, Stanford University.

Hein, G.E. (1987). The right test for hands-on learning? Science and Children, 25(2), 8-12.

Linn, R.L. (1992). Achievement testing. In M.C. Alkin (Ed.) Encyclopedia of Educational Research (pp 1-12). New York: Macmillan Publishing Company.

National Research Council (1995). National science education standards. Washington, DC: Author.

Raisen, S.A., Baron, J.B., Champagne, A.B., Haertel, E. Mullis, I.V.S., Oakes, J. (1990). Assessment in science education: The middle years. Andover, MA: The National Center for Improving Science Education.

Shavelson, R.J., & Ruiz-Primo, M.A. (in press). On the assessment of Science Achievement. Unterrichtswissenschaft.

Shavelson, R.J., Solano-Flores, G., & Ruiz-Primo, M.A. (in press). Toward a science performance assessment technology. Educational Evaluation and Planning.

Snow, R.E. (1968). Brunswikian approaches to research on teaching. American Educational Research Journal, 5, 475-489.

Solano-Flores, G., Shavelson, R.J., Ruiz-Primo, M.A., Schultz, S.E., Wiley, E., & Brown, J. (1997, March). On the development and scoring of classification and observation science performance assessments. Paper presented at AERA annual Meeting. Chicago, IL.

Solano-Flores, G., & Shavelson, R.J. (1997). Development of performance assessments in science: conceptual, practical, and logistical issues. Educational Measurement: Issues and practices, 16(3), 16-25.

Stecher, B.M. & Klein, S.P. (1995). Performance assessment in science: Hands-on tasks and scoring guides (Technical Report DRU-1087-NSF). Santa Monica, CA: RAND Corporation.

Wolf, R.M. (1990). A framework for evaluation. In H.J. Walberg & G.D. Haertel (Eds.) The international encyclopedia of educational evaluation (pp 61-66). Oxford: Pergamon Press.

Appendix A

Raw Scores Descriptive Statistics by Class and Sequence Across the Two Units

Sequence	Class	Variables							Mixtures						
			Pretest			Postest				Pretest			Postest		
		n	Max	Mean	S.D.	Max	Mean	S.D.	n	Max	Mean	S.D.	Max	Mean	S.D.
Close-Close	1	5	16	10.20	1.52	16	12.00	.63	5	20	4.00	3.08	20	6.70	4.51
	2	5		5.60	5.47		8.40	2.19	5		9.70	3.46		9.50	1.32
	3	6		7.75	2.75		9.00	1.83	6		2.50	1.61		6.58	4.53
	4								7		5.79	4.33		9.92	3.98
	All	16		7.84	.96		9.75	1.00	23		5.39	.86		8.26	.82
Proximal-Proximal	1	7	29	15.93	5.74	29	16.43	2.20	6	18	9.79	4.46	18	9.29	3.93
	2	4		13.25	5.45		11.88	1.20	8		9.16	1.78		8.25	4.27
	3	6		17.50	4.94		18.83	1.75	5		3.40	3.13		3.20	3.27
	4								7		10.61	3.48		12.21	2.12
	All	17		15.85	1.29		16.21	1.26	26		8.59	.79		8.59	4.54
									5						
Close-Proximal	1	7	16	11.07	4.01	29	17.50	1.35	6	20	3.00	1.70	18	9.54	3.39
	2	5		4.60	4.62		11.30	2.11	7		6.57	3.45		9.57	3.55
	3	7		10.5	3.29		19.36	2.93	5		3.10	1.52		7.40	4.34
	4								6		4.33	3.15		9.63	4.50
	All	19		9.16	1.07		16.55	1.45	24		4.40	.93		9.13	3.78
Proximal-Close	1	7	29	16.79	4.86	16	10.21	.29	6	18	11.04	2.25	20	9.83	3.19
	2	4		12.87	4.59		7.25	1.90	4		7.87	2.46		5.38	3.82
	3	7		15.64	4.96		13.86	.51	3		.66	1.15		1.50	1.32
	4								7		10.11	2.62		11.29	4.56
	All	18		15.47	1.30		10.97	.76	20		8.53	.93		8.20	1.11